Title
Discriminant Diagnostics

Author
Peter A. Lachenbruch

Publication Date
2011-10-24

BIOMETRICS 53, 1284–1292
December 1997

Discriminant Diagnostics

Peter A. Lachenbruch
FDA/CBER/OELPS/DBE, 1401 Rockville Pike, HFM-215,
Rockville, Maryland 20852, U.S.A.

SUMMARY

I discuss diagnostic methods for discriminant analysis. The equivalence with linear regression is noted and regression diagnostics are considered. The leverage is a function of the linear discriminant function and the Mahalanobis distance of the observation from the group mean. The distribution of this distance is approximately chi-square with degrees of freedom equal to the number of variables. Standard tests may be used to examine the normality assumption. Some examples are given.

1. Introduction

An important component of a regression analysis is the investigation of data points that may distort the analysis because of high influence or unusual values of the dependent or independent variables (Cook and Weisberg, 1982). This investigation may result in removing the unusual observation from the analysis or in reducing its weight. Similar concerns exist in two-group discriminant analysis because an observation may be unusual, the discriminant coefficients may be affected, and the analysis rendered suspect. We want to detect the observations that cause large changes in the discriminant coefficients. In regression, the values of the independent variables are (in theory, at least) under the control of the investigator. In discriminant analysis, all observations other than the group identifier are presumed to be random variables.

Using group indicators as the dependent variable in a multiple regression leads to a set of coefficients that are proportional to the linear discriminant function (LDF). A simple way to develop discriminant diagnostics is to consider the usual regression diagnostics applied to this multiple regression. This has some disadvantages: the diagnostics are not directly interpretable in terms of the discriminant function, and for two major diagnostics, we obtain no information. The plot of the residuals against the estimated values will plot $y - \mathbf{b}'\mathbf{x}$ against $\mathbf{b}'\mathbf{x}$. Here, y is 1 or 0, so the plot must be two straight lines with slope -1. This provides no information about possible outlying points or variance heterogeneity, which this plot is designed to detect. Similarly, studentized residuals have almost the same flaw: we obtain very little information, as the plot is generally close to the two straight lines with slope -1. The leverage statistics from the regression model are interpretable as functions of the distance of the point from its group mean and of its discriminant score.

Campbell (1978) derived the influence function for general discriminant functions and Mahalanobis distances. He found that the influence was a (messy) function of the discriminant score and the Mahalanobis distance between groups. Some of his work is similar to the results given below. Critchley and Vitiello (1991) examined the influence of observations on misclassification and confirmed results of Campbell with an exact expression. They noted that the influence is governed by the difference between the linear discriminant score and by how atypical it is in its own population. They note the similarity to the leverage. In this note, we show how the leverage is a function of these two quantities. Fung (1995) has also discussed this and noted that Cook's distance is a function of these quantities. He also notes that the interpretation of these measures has a fundamental difference in regression and discrimination. Since the $\mathbf{X}$ values are assumed to be fixed, in regression, the leverage is a multivariate outlier measure, while for the discriminant context the $\mathbf{X}$ values are random variables and such an interpretation is not appropriate.

Key words: Diagnostics; Discriminant function; Leverage; Mahalanobis distance.

We shall consider the two-group discriminant problem in which the groups have a common covariance matrix. The usual LDF arises if we assume multivariate normality of the observations $\mathbf{X}$ or if we find the linear function that maximizes the ratio of the between-groups sum of squares to the within-group sum of squares (see, e.g., Lachenbruch, 1975). The normality assumption can be used when we consider diagnostic plots. In the following, we will let

$$D_s(\mathbf{X}) = (\mathbf{X} - 1/2(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2))'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

be the sample discriminant function, where $\mathbf{S}$ is the $(k \times k)$ sample pooled covariance matrix and $\bar{\mathbf{x}}_i$ is the $(k \times 1)$ sample mean of the ith population. $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is the Mahalanobis distance between groups.

A simple statistic to consider is the LDF itself, standardized to zero mean and unit variance. $T(\mathbf{X}) = (D_s(\mathbf{X}) - D_s(\bar{\mathbf{x}}_i))/(\text{var}(D_s(\mathbf{X})))^{1/2}$ is easily seen to be proportional to $(\mathbf{X} - \bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. This will be approximately standard normal if the observations are multivariate normal. The standardization is convenient for detecting outlying observations; it is not crucial to the discussion. It also appears to be fairly robust to nonnormality of the observations, although relatively little is known for low dimensions (Lachenbruch, 1975). Most of the robustness research has been done on five or more variables. The population means can be estimated by the corresponding sample means. Normal plots and general outlier examination can be applied to $T(\mathbf{X})$.

There are several useful identities that we use. We derive statistics for observations from the first population. All results carry over to the second. The ith deleted mean is the mean of the observations after deleting the ith observation and is denoted as $\bar{\mathbf{x}}_{1(i)}$. The first identity gives $\bar{\mathbf{x}}_1 = \bar{\mathbf{x}}_{1(i)} + \mathbf{d}_{1(i)}/n_1$, where $\mathbf{d}_{1(i)} = \mathbf{x}_{1i} - \bar{\mathbf{x}}_{1(i)}$. We also have the relation $\bar{\mathbf{x}}_{1(i)} = \bar{\mathbf{x}}_1 - \mathbf{d}_{1i}/(n_1 - 1)$, where $\mathbf{d}_{1i} = \mathbf{x}_{1i} - \bar{\mathbf{x}}_1$ and n_1 is the number of observations in the first group. The second identity states that, if we can write a matrix as $\mathbf{A} + \mathbf{u}\mathbf{v}'$, where $\mathbf{u}$ and $\mathbf{v}$ are column vectors, then

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - (\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1})/(1 + \mathbf{u}'\mathbf{A}^{-1}\mathbf{v}).$$

Each of these can be verified algebraically. The second identity is useful when computing statistics excluding the ith observation.

2. Leverage

We shall develop diagnostic statistics for observations from the first group. The dependent variable, y, has values 1 in group 1 and 0 in group 2. A referee has noted that, in the usual framework, the dummy variable takes the values $n_2/(n_1+n_2)$ in group 1 and $-n_1/(n_1+n_2)$ in group 2. This changes the intercept, but has no effect on the coefficients. The spacing between the values of y is the same. Denote the overall mean by $\bar{\mathbf{x}}$, which is $k \times 1$. Let $\mathbf{J}$ be an $n \times 1$ vector of 1's. Then $\mathbf{J}\bar{\mathbf{x}}' = \mathbf{m}$ is the $n \times k$ matrix of the grand mean and the vector of differences of the observations from this is $\mathbf{x} - \mathbf{m}$ (note that this is an $n \times k$ matrix). The regression coefficients are $((\mathbf{x} - \mathbf{m})'(\mathbf{x} - \mathbf{m}))^{-1}(\mathbf{x} - \mathbf{m})'\mathbf{y}$. We can show that

$$\begin{aligned} \mathbf{V} &= ((\mathbf{x} - \mathbf{m})'(\mathbf{x} - \mathbf{m})) \\ &= (n_1 + n_2 - 2)\mathbf{S} + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'n_1 n_2/(n_1 + n_2) \\ &= c_1\mathbf{S} + c_2(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)', \end{aligned}$$

where $c_1 = n_1 + n_2 - 2$ and $c = n_1 n_2/(n_1 + n_2)$.

Let

$$\mathbf{d}_i = \mathbf{x}_i - \bar{\mathbf{x}} = \mathbf{x}_i - \bar{\mathbf{x}}_1 + c_3(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \mathbf{d}_i + c_3(\bar{\mathbf{x}} - \bar{\mathbf{x}}_2),$$

where $c_3 = n_2/(n_1 + n_2)$. Using the identity from Section 1, let $\mathbf{a} = c_1\mathbf{S}$ and $\mathbf{u} = \mathbf{v} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'(c_3)^{1/2}$. The leverage is defined as the diagonal elements of $\mathbf{H} = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{V}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})'$. For the ith point in group 1, the leverage is

$$\mathbf{d}_{1i}'\mathbf{V}^{-1}\mathbf{d}_{1i} = (\mathbf{d}_i + c_3(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))'\mathbf{V}^{-1}(\mathbf{d}_i + c_3(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)),$$

which is

$$\begin{aligned} h_i = \{1/c_1[&\mathbf{d}_i'\mathbf{S}^{-1}\mathbf{d}_i - c_2/c_1(\mathbf{d}_i'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2/(1 + (c_2/c_1)D^2) \\ &+ 2c_3\mathbf{d}_i'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - 2(c_2c_3/c_1)\mathbf{d}_i'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)D^2/(1 + (c_2/c_1)D^2) \\ &+ c_3^2(D^2 - (c_2/c_1)D^4/(1 + (c_2/c_1)D^2))]\}. \end{aligned}$$

The first term is the Mahalanobis distance $\phi_i = \mathbf{d}_i'\mathbf{S}^{-1}\mathbf{d}_i$ of the observation from the sample mean, the term $\mathbf{d}_i'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D_s(\bar{\mathbf{x}}_i) - D_s(\bar{\mathbf{x}}_1) = L_i$ is the difference of the LDF at $\mathbf{x}_i$ and the LDF at $\bar{\mathbf{x}}_1$. With some algebra, the leverage can be expressed as

$$h_i = 1/c_1[\phi_i - \{c_2/c_1 L_i^2 - 2c_3 L_i - c_3^2 D^2\}/(1 + (c_2/c_1)D^2)],$$

which is a function of the Mahalanobis distance ϕ_i, the discriminant distance of the discriminant function of the point from the discriminant function of the group mean, the sample sizes, and the Mahalanobis distance between the groups, D^2. Note that L_i is the sample estimate of the numerator of the within-group standardized discriminant statistics mentioned in the Introduction. The only quantities that are a function of the ith observation are ϕ_i and L_i. Leverage is an increasing function of ϕ_i and a decreasing function of L_i. It is unclear how the two terms behave for observations near the mean. Clearly, ϕ_i and L_i are related. An informal argument is that

$$L_i^2 = \mathbf{d}_i'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\mathbf{d}_i \approx \mathbf{d}_i\mathbf{S}^{-1}\mathbf{d}_i/c_2 = \phi_i/c_2$$

since the covariance matrix of $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ is Σ/c_2.

If $\phi = 0$, then $L = 0$ (since then $\mathbf{x} = \bar{\mathbf{x}}_1$). For $\mathbf{x}$ far from $\bar{\mathbf{x}}_1$, the influence will be large since ϕ is a quadratic function of $\mathbf{d}$. Note that L_i is weighted inversely by D^2, so for well separated groups, the contribution of L to the leverage will be somewhat less than ϕ.

Since $\mathbf{x} - \bar{\mathbf{x}}_1$ is $\mathrm{MVN}(\mathbf{0}, \Sigma(n_1 - 1)/n_1)$ if $\mathbf{x}$ is not an outlier, $n_1/(n_1 - 1)\phi$ is a T^2 variable, which is proportional to an F. Penny (1996) notes that the usual criterion $p(n_1 - 1)F/(n_1 - p)$ is too small and suggests that it should be replaced by $p(n_1 - 1)^2 F'/[n_1(n_1 - p - 1 + pF')]$, where $F' = F_{p,n-p-1,\alpha/n}$. This provides a Bonferroni adjustment for examining each of the observations. For large values of $n = n_1 + n_2$, this is about pF'', which is approximately a χ^2 with degrees of freedom equal to the number of variables in the LDF. This suggests that a plot of the ordered values of ϕ_i against the percentiles of χ^2 will be helpful in examining this diagnostic. Large values of ϕ will be at the positive extreme of the χ^2 distribution. Using the Bonferroni adjustment, for two variables, the 90th percentile is 14.06 and for four variables it is 18.74. Penny's criteria gives 13.32 and 17.53. These suggest that the χ^2 values are too large and will not detect outliers often enough. The Bonferroni adjustment is fairly conservative, so I have used $\alpha = 0.9$ rather than $\alpha = 0.95$ or 0.99. Values of ϕ greater than the $(100 - 10/n)$th percentile would cause me to examine such a point closely. Similarly, L_i is a shifted discriminant function and is approximately normal with mean 0 and variance approximately D^2 (more precise statements could be made, but are not useful for graphical procedures). Thus, if $|L|$ is larger than $3.13D$, we should examine the point further (this adjusts for the multiple tests). Note that these Bonferroni adjustments are for the full data set. A plausible argument might be made for applying the adjustment to each group separately. This would make it easier to reject observations.

If the output provides values of ϕ from each group, one should examine these values within groups. The examination might consist of comparing the value of ϕ to the appropriate critical value of χ^2 or a graphical examination of a plot against expected values from a χ^2 distribution. If a value is implausibly large, some action is appropriate. The actions available include deleting the observation, downweighting the observation, and shrinking the observation along the ray between it and the group mean. If the original data sheets are available, they should be checked to ensure that the data values have been correctly entered. Examining a normal plot of L for each group (or the discriminant function) will provide information on the applicability of normal theory. A histogram of the leverage will give us similar information. It is a mixture of contributions from ϕ and L. The correlation of ϕ and the leverage is high in the examples I've examined. The leverage is a good reflection of the Mahalanobis distance.

3. Examples

First, let us examine the behavior of ϕ and L for some values of an observation $\mathbf{X}$ from group 1. We consider observations at $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$, and at the pooled mean. For these points, we have

	$\mathbf{X}$		
Statistic	$\bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_2$	$(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$
ϕ	0	D^2	$D^2/4$
L	0	$-D^2$	$-D^2/2$

Table 1
Means and covariances for shock data

Group (n)	SP	MAP	MVP	PVI	UO
			Means		
Survived (70)	114.60	79.70	7.80	49.50	78.60
Died (43)	92.30	63.20	10.70	47.50	15.20
			Covariances		
SP	830.61				
MAP	526.54	422.68			
MVP	11.04	2.02	30.66		
PVI	−49.11	−66.03	14.55	234.19	
UO	145.74	130.34	−107.56	51.18	11,771.51

Thus, if the point is near the mean of the alternative group, both ϕ and L will be large. A point near the group mean will have a small ϕ and L. A value of $D^2/4$ must be interpreted as a function of the number of variables in the discriminant function. For example, if $D^2 = 25$ and there are two variables, we would be quite concerned. If there were 10 variables, we would not be. A outlying point orthogonal to the plane defined by the discriminant coefficients $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ will have larger ϕ than L. L will be sensitive to outliers with large or small discriminant scores.

We use the shock data set from Afifi and Azen (1972). Data were collected on many patients admitted to the hospital for the treatment of shock. We use the variables survival status (SURV), systolic pressure (SP), mean arterial pressure (MAP), mean venous pressure (MVP), plasma volume index (PVI), and urinary output (UO). These variables show substantial differences in the two groups on mean arterial pressure, mean venous pressure, and urinary output, but not on systolic blood pressure or plasma volume index. The two groups are those who survived and those who died. The authors give measurements for patients made at admission and prior to discharge or death; we use only the initial measurement. For these data, the means and within-group covariance matrix are given in Table 1.

Figures 1a and 1b give the scatter-plot matrices of these data. One sees that there are several suspicious points. These outliers may be characterized as being bivariate outliers with a few exceptions: there is one very large MVP value and a few large UO values. Figures 2a and 2b give normal plots of these with the appropriate line shown in the plot. It is immediately evident that UO is far from normal. A large proportion of these patients have a UO of 0, leading to the obviously nonnormal plots. The slopes of the plots appear to be the same, suggesting that the variances are the same (this says nothing about equal correlations, however). The Kolmogorov–Smirnov test with the Lilliefors adjustment indicates that UO is clearly nonnormal in both groups and that the PVI may be slightly off in the survival group ($p = 0.02$) but not in the death group. There is no cause for concern in the other variables.

The discriminant function (from SYSTAT output) using all variables is

$$D_s(\mathbf{X}) = 0.0102 * \mathrm{SP} + 0.0297 * \mathrm{MAP} - 0.1010 * \mathrm{MVP} + 0.0246 * \mathrm{PVI} + 0.0039 * \mathrm{UO} - 0.6586.$$

With this function, we allocate 48 of the 70 survivors correctly and 32 of the 43 deaths correctly. The discriminant scores range from -1.7 to 2.8 in the survivor group and from -2.7 to 1.5 in the death group, with a considerable overlap. The D^2 is 1.3172 (computed from $F = 6.7645$ with 5 and 107 d.f.). We used the 99.91st percentile of χ^2 as a criterion for flagging suspect observations using ϕ. For these data, $\chi^2(5, 0.9991) = 20.80$. No patients would be judged to be atypical of their groups using this criterion. Penny's critical value, 19.37, suggests that case 29 is suspect. If we use the 99th percentile of $\chi^2 (= 15.08)$, we would reject more frequently. To check for extreme values of the discriminant score, we use L. In group 1, since $L = D_S(\mathbf{X}) - D^2/2$, the values that concern us are those discriminant scores greater than 4.905 or less than -2.929 (using the 99.91st percentile of $z = 3.127$). The negatives of these values may be used in group 2 (this uses the plug-in estimate of the standard deviation of L, $\mathrm{var}(L) = D = 1.148$). There were no values of L that suggested problems in either group. The values of the variables for the suspicious (and near suspicious) cases were (ϕ_1 and ϕ_2 are the distances from the centroids of groups 1 and 2):

Case	SP	MAP	MVP	PVI	UO	ϕ_1	ϕ_2	Leverage
15	113	84	4.0	39.3	510	16.6	21.8	0.164
29	171	117	30.2	49.4	0	19.7	19.2	0.182
57	82	61	15.2	106.6	1	16.9	17.2	0.159

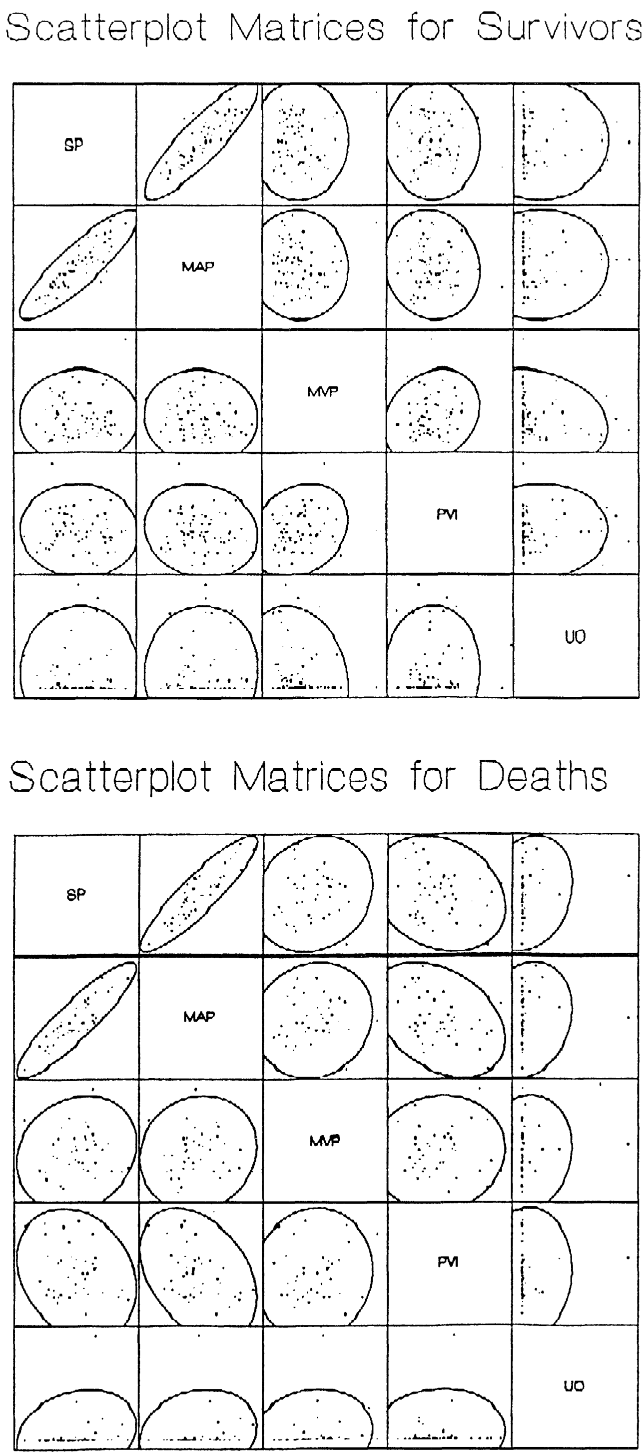


Figure 1. **a.** Scatterplot matrix for survivors. **b.** Scatterplot matrix for deaths.

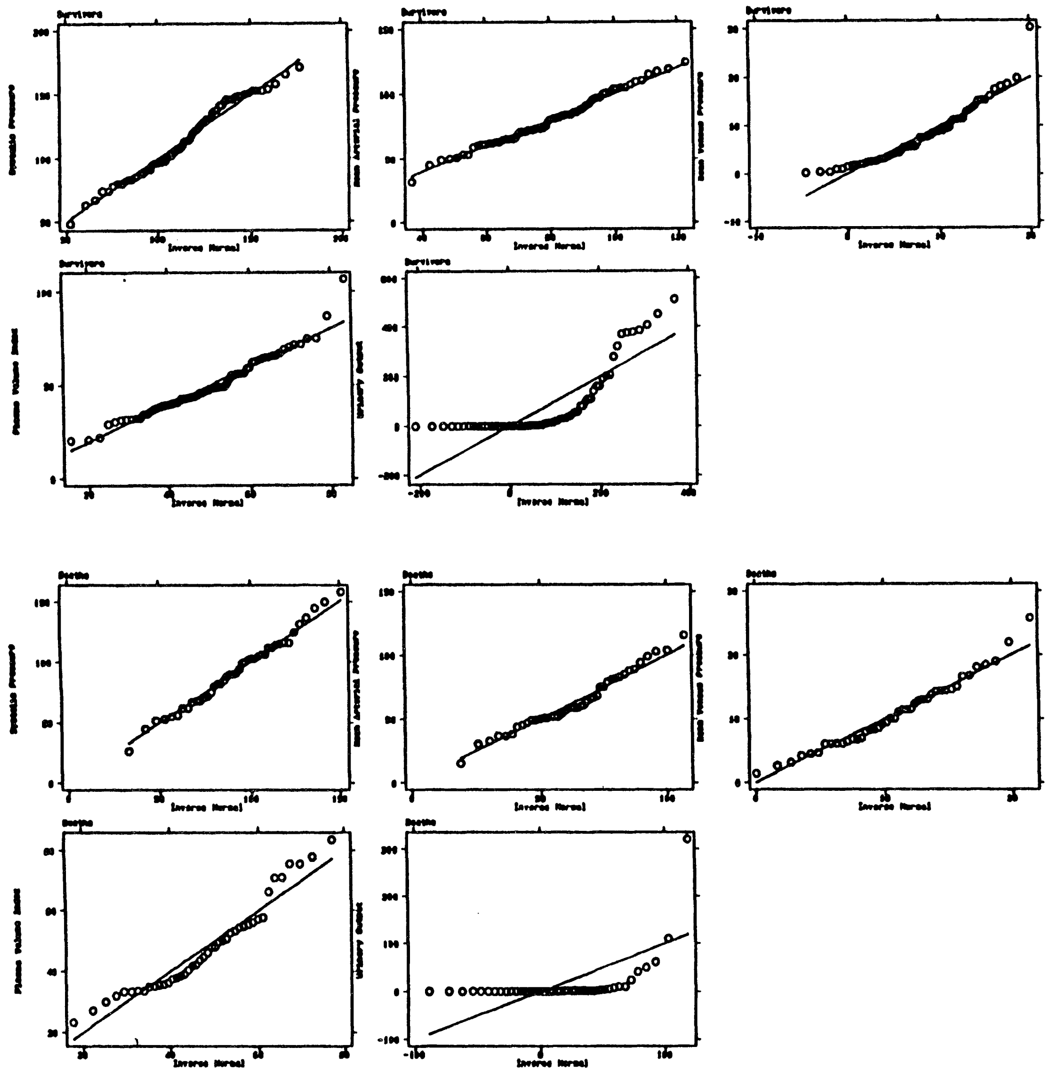


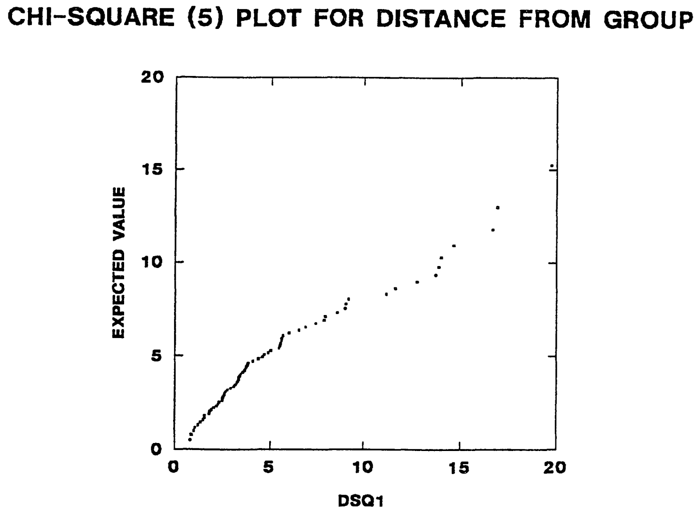
Figure 2. **a.** Normal probability plots of predictor variables for survivors. **b.** Normal probability plots of predictor variables for deaths.

These cases have values indicating they are unlike the survivors. Case 15 has the largest UO value among the survivors. Case 29 has a very low UO value and the highest SP among the survivors. Case 57 has a low UO value and the highest PVI value among the survivors. These cases are so atypical of either group of patients that they might best be deleted from the data set. They would be deleted using the more liberal 0.99 value. After removing them from the data, the discriminant function becomes

$$D_s(\mathbf{X}) = 0.0116 * \mathrm{SP} + 0.0245 * \mathrm{MAP} - 0.1216 * \mathrm{MVP} + 0.0185 * \mathrm{PVI} + 0.0041 * \mathrm{UO} - 0.6669.$$

We note that the coefficient of MAP declined by 21%, that of MVP by 17%, while that of PVI increased by 33%. Thus, these three observations had a major effect on the discriminant coefficients. Figures 3a and 3b give the χ^2 plots for 5 d.f. for survivors and deaths. In the survivors, the three suspicious points at the upper tail correspond to the points with large values shown above. In the plot for deaths, one point appears to be large, but the scales of the two diagrams are different. Its value is less than the comparison value (15.08) and was not further studied.

The assumption of normality of the discriminant function is often used to justify a plug-in estimate of the error rate $\Phi(-D/2)$ (assuming equal prior probabilities). Since at least one of the variables in the discriminant function is clearly not normal, it is appropriate to examine the com-



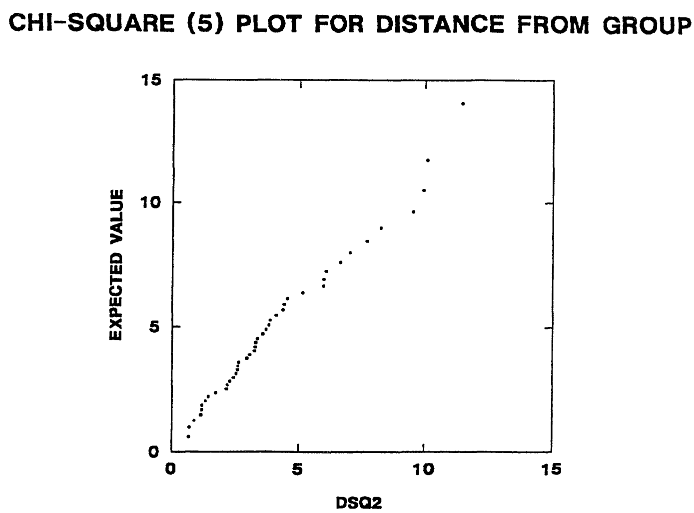
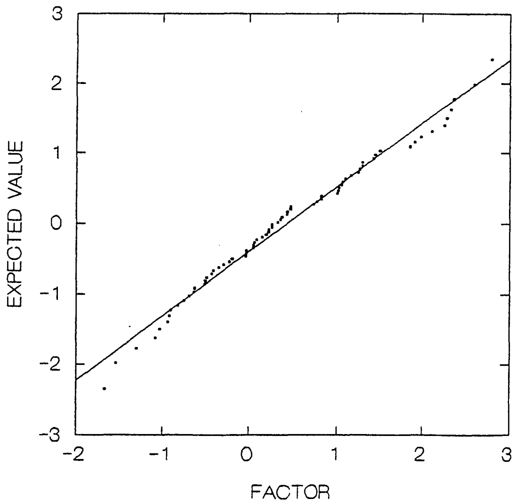


Figure 3. **a.** Chi-square (5) probability plot for distance from group mean for suvivors. **b.** Chi-square (5) probability plot for distance from group mean for deaths.

puted discriminant functions for normality. Figures 4a and 4b give normal plots of $D_s(\mathbf{X})$ for the two groups. They appear to be satisfactory. We also used the Kolmogorov–Smirnov test with the Lilliefors adjustment (Wilkinson, 1995) and did not find evidence for non-normality in either group. This supports general robustness findings about the LDF (e.g., Lachenbruch, 1975).

4. Discussion

The question of the appropriate alarm level for the outlier statistics naturally arises. Since their distributions are approximately known, rough critical values may be obtained from the χ^2 distribution for ϕ and for the normal distribution for L. Using conservative critical values is appropriate, in my opinion, since we don't know which observations are suspect *a priori* and multiple tests will be done. In the example, we deleted the three observations with large values using the 0.99 critical value. Penny's approach (1996) using the $1 - \alpha n$ $(= 0.9991)$ critical value would delete a single observation. These are arbitrary; a third alternative would be to shrink the data along the ray from the group mean. Broffitt, Clarke, and Lachenbruch (1980) suggested this. They obtained good results in a scale contamination problem by forming the pseudo-observation along that ray with length $\min(\phi, K)$, where K was a suitably chosen percentile of D^2. Because L will be normally distributed if $\mathbf{X}$ is, it is useful in detecting deviations from normality of the discriminant function. Since L is a weighted sum of observations, it is possible that it will appear normal even when some



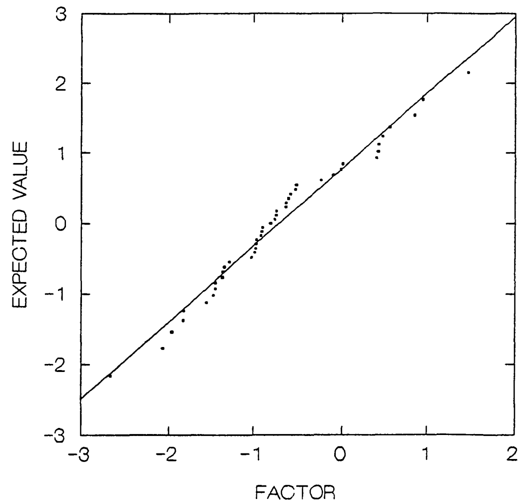


Figure 4. **a.** Normal probability plot of discriminant function for survivors. **b.** Normal probability plot of discriminant function for deaths.

of its components are not. In the Afifi/Azen data, no lack of normality of the discriminant function was found, despite the gross lack of normality in the UO (urinary output) variable.

Unequal covariance matrices may not be detected with these statistics. There are general multivariate statistics for this purpose (see, e.g., Anderson, 1984). Unfortunately, they are affected by violations of the multivariate normality assumption.

Wilks (1963) considered tests for multivariate outliers based on the likelihood ratio statistic. The Mahalanobis distance statistic is equivalent to this. A referee notes that "the critical values given by the χ^2 approximation [for the Mahalanobis distance] can be very different from those based on Wilks."

The computation of these statistics is a straightforward procedure in several statistical packages. BMDP (Dixon, 1994) gives the Mahalanobis distance ϕ of each observation from both group means as part of its standard output. Thus, it is easy to examine the data for outliers. SYSTAT (Wilkinson, 1995) provides a method of saving output in the MGLH module, which gives from each group as part of its output. SYSTAT also will produce plot values against a χ^2 distribution. SAS (SAS Institute, Inc., 1994) does not give the Mahalanobis distance as part of standard output. However, it is easy to write a SAS IML program to provide this. STATA (Stata, 1995) does not provide this as output, but one can use its matrix language to give this information.

A general diagnostic and prevention procedure might be as follows. Before computing the discriminant function, examine the individual variables for nonnormality and transform when possible. After computing the discriminant function, use the leverage or the Mahalanobis distance (if it's given) to determine if some points are outliers and should be deleted or downweighted. If a point is extremely far from both groups, it may be an indication of an additional group that has not been suspected or has been deliberately incorporated with one of the groups. An example of this could be in the case when a nondiseased group is being contrasted to a group consisting of several related diseases (say of a particular organ). This would suggest that the presumed homogeneity of the groups did not, in fact, hold.

Acknowledgements

This work was done while the author was at the UCLA Department of Biostatistics. The author thanks the referees and the associate editor for their perceptive review of this paper. In particular, the expression for the leverage is due to the referee. The views expressed here are those of the author and are not necessarily those of the U.S. Food and Drug Administration.

Résumé

L'auteur étudie le problème de l'influence des observations en analyse discriminante classique. Il rappelle l'étroite relation qui existe entre l'analyse discriminante et la régression multiple et il profite des méthodes diagnostiques qui ont été développées dans ce domaine. Il montre que la fonction d'influence (leverage) dépend de la fonction linéaire discriminante de Fisher et de la distance de Mahalanobis entre une observation et la moyenne de son groupe. Cette distance est approximativement distribuée comme un chi-carré dont le nombre de degrés de liberté est égal au nombre de variables du problème. Les hypothèses de normalité peuvent être testées à l'aide de statistiques classiques. Des exemples illustrent les nouvelles méthodes proposées.

References

Afifi, A. and Azen, S. (1972). *Statistical Analysis: A Computer Oriented Approach.* New York: Academic Press.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis,* 2nd edition. New York: John Wiley and Sons.

Broffitt, B., Clarke, W. R., and Lachenbruch, P. A. (1980). The effect of huberizing and trimming on the quadratic discriminant function. *Communications in Statistics: Theory and Methods* **A9**(1), 13–25.

Campbell, N. (1978). The influence function as an aid in outlier detection in discriminant analysis. *Applied Statistics* **27**(3), 251–258.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Critchley, F. and Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika* **78**(3), 677–690.

Dixon, W. J. (chief ed). (1994). *BMDP Statistical Software Manual.* Berkeley: University of California Press.

Fung, W. K. (1995). Diagnostics in linear discriminant analysis. *Journal of the American Statistical Association* **90**(431), 952–956.

Lachenbruch, P. A. (1975). *Discriminant Analysis.* New York: Hafner Press.

MacLachlan, G. J. (1991). *Discriminant Analysis and Statistical Pattern Recognition.* New York: John Wiley and Sons

Penny, K. I. (1996). Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Applied Statistics* **45**(1), 73–81.

SAS Institute, Inc. (1994). *SAS User's Guide: Statistics,* Version 6.10 edition. Cary, North Carolina: SAS Institute.

Stata Corporation. (1994). *Stata Reference Manual,* Release 4.0. College Station, Texas: Stata.

Wilkinson, L. (1995). *SYSTAT: The System for Statistics.* Chicago: SYSTAT.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhya, Series A* **25,** 407–426.

Received May 1994; *revised June* 1995 *and January* 1997; *accepted April* 1997.